Two Oracle Inequalities for Regularized Boosting Classifiers

Ingo Steinwart
Information Sciences Group CCS-3
Los Alamos National Laboratory
Los Alamos, NM 87545, USA
ingo@lanl.gov

November 5, 2008

Abstract

We derive two oracle inequalities for regularized boosting algorithms for classification. The first oracle inequality generalizes and refines a result from [4], while the second oracle inequality leads to faster learning rates than those of [4] whenever the set of weak learners does not perfectly approximate the target function. The techniques leading to the second oracle inequality are based on the well-known approach of adding some artificial noise to the labeling process.

1 Introduction

One often employed method of finding a classifier with the help of empirical data $D := ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$ is that of regularized boosting, see e.g., [11]. In this approach, a family $(e_i)_{i \in I}$ of weak classifiers $e_i : X \to \mathbb{R}$ is given, and, with the help of D, a weighted combination $f_{w^*} := \sum_{i \in I} w_i^* e_i$ is constructed, where $w^* := (w_i^*)_{i \in I}$ is a real-valued family that satisfies

$$\lambda \sum_{i \in I} |w_i^*| + \frac{1}{n} \sum_{i=1}^{n} L(y_i, f_{w^*}(x_i)) < \inf_{w} \lambda \sum_{i \in I} |w_i| + \frac{1}{n} \sum_{i=1}^{n} L(y_i, f_w(x_i)) + \varepsilon \quad (1)$$

for some regularization parameter $\lambda > 0$, some convex loss function L, e.g. the logistic loss for classification, and some numeric tolerance $\epsilon \geq 0$. Here the family of weak classifiers may, e.g., be the output of some classification algorithms such as neural nets, decision trees, or support vector machines. In this case, boosting may be viewed as an alternative to the often used parameter selection step required by these algorithms. However, the family of weak classifiers may also be a family of particular simple functions such as decision stumps that are not output of a previous classification algorithm. We refer to [11] for more information in this regard. Moreover, recall that the regularization term was motivated by the fact the early boosting methods such as

AdaBoost may overfit in the presence of label noise, see, e.g., again [11] and the references therein. Another approach to resolve this potential overfitting is early stopping, which has been discussed by [18, 2]

In recent years boosting algorithms have been successfully applied in various application areas, such as optical character recognition, natural language processing, face recognition, cancer detection, and text classification. We refer again to the survey [11] for more applications and corresponding references. In this regard we note that a particular nice feature of boosting algorithms is that basically no assumptions on the family of weak classifiers need to be made. In particular, the input space X is not required to be a subset of $\mathbb{R}^d$, which opens, like for support vector machines, the possibility to deal with non standard data formats. For support vector machines, however, this flexibility is only possible, if a reasonable kernel on X is available, which, at least in some circumstances, may be not the case. In contrast to this, the boosting algorithm (1) does not need such requirements on its base function class determined by the family of weak learners, and hence it may be applicable in potentially more situations. Last but not least, the optimization problem (1) is convex in w, and hence regularized boosting offers computational properties similar to those of support vector machines. We refer yet another time to [11] for a detailed list of references. Finally, some more information on boosting, which complements [11], can be found in [8].

For boosting methods based on optimization problems related to (1), the articles [10, 4, 1, 18] establish both universal consistency and learning rates, where [4] considers an algorithm that, up to a discretization and some minor technical details, resembles (1). So far, however, consistency and learning rates for the original approach described by (1) have not been established. The first goal of this work is to close this gap by establishing an oracle inequality for regularized boosting based on (1). From this oracle inequality, we then derive universal consistency and learning rates under natural assumptions on the family $(e_i)_{i\in I}$ and the data-generating distribution P, where the learning rates match those of [4] for the discretized version of (1). As already observed in [4], these learning rates become better for the logistic loss, if the posterior probability η of P is bounded away from the levels 0 and 1, i.e., if there is noise in each label. As a consequence, [4] suggested to add some artificial noise to the labels. Our second goal of this work is to establish an oracle inequality for this approach. Here it turns out that, if the family of weak classifiers approximates the target function in an optimal way, see Lemma 2.3 for a precise statement of optimality, then this new oracle inequality leads to the same learning rates as our first oracle inequality does. In the absence of such perfect approximation, however, the new oracle inequality *always* leads to faster learning rates. Note that this better behavior is of particular interest, if the used weak classifiers are the output of a classification algorithm, since in this case perfect approximation can almost never be guaranteed.

The rest of this work is organized as follows. In Section 2 we introduce all necessary concepts, present our two oracle inequalities, and discuss some of their consequences including consistency and learning rates. Section 3 contains all proofs.

2 Main results

In the first part of this section, we introduce all necessary notions for presenting our main results. Our first oracle inequality is then presented and discussed in Subsection 2.2, while the second oracle inequality is considered in Subsection 2.3.

2.1 Preliminaries

In the following we always write $Y := \{-1, 1\}$. Moreover, X always denotes a complete measurable space and P a distribution on $X \times Y$. We call a measurable function $L : Y \times \mathbb{R} \to [0, \infty)$ a loss, and if there exists a $\varphi : \mathbb{R} \to [0, \infty)$ such that

$$L(y, t) = \varphi(yt)\,, \qquad y \in Y,\, t \in \mathbb{R},$$

we say that L is a margin-based loss. In this case, we call φ the representing function of L. Various loss functions used in classification algorithms are margin-based, here we only mention the hinge loss, the (truncated) least squares loss, the logistic loss represented by $\varphi(t) := \ln(1 + \exp(-t))$, $t \in \mathbb{R}$, and the AdaBoost loss represented by $\varphi(t) := \exp(-t)$, $t \in \mathbb{R}$. For some simple properties of these losses we refer to [1] and [14, Chapter 2.3]. Moreover, we need the classification loss $L_{\mathrm{class}} : Y \times \mathbb{R} \to [0, \infty)$ defined by

$$L_{\mathrm{class}}(y, t) := \mathbf{1}_{(-\infty,0]}(y \operatorname{sign} t)\,, \qquad y \in Y,\, t \in \mathbb{R},$$

where $\operatorname{sign} 0 := 1$ and $\mathbf{1}_A$ denotes the indicator function of a set A. Clearly, L_{class}, which is used to define the learning goal of binary classification, is *not* margin-based.

In the following, we say that a loss L is (strictly) convex or continuous, if and only if $L(y, \cdot) : \mathbb{R} \to [0, \infty)$ is (strictly) convex or continuous for all $y \in Y$, respectively. While all the margin-based losses considered above are both convex and continuous, L_{class} does not satisfy either of these properties. Furthermore, we say that a loss L is locally Lipschitz-continuous if for all $a \geq 0$ there exists a constant $c_a \geq 0$ such that

$$\bigl|L(y, t) - L(y, t')\bigr| \,\leq\, c_a\,|t - t'|\,, \qquad y \in Y\,, t, t' \in [-a, a]\,.$$

Moreover, for $a \geq 0$, the smallest such constant c_a is denoted by $|L|_{a,1}$. Finally, if we have $|L|_1 := \sup_{a \geq 0} |L|_{a,1} < \infty$, we call L Lipschitz continuous. For margin-based losses, we refer to [14, Lemma 2.25] for some simple connections between these notions. In particular, recall that convex, margin-based losses are always locally Lipschitz-continuous. Finally, we say that a margin-based loss L is k-times continuously differentiable, if its representing function φ is k-times continuously differentiable.

Given a loss function L and a function $f : X \to \mathbb{R}$, we often write $L \circ f$ for the function $X \times Y \to [0, \infty)$ defined by

$$L \circ f(x, y) := L(y, f(x))\,, \qquad y \in Y,\, x \in X\,.$$

Now let P be a distribution on $X \times Y$. For a loss $L : Y \times \mathbb{R} \to [0, \infty)$ we then define the L-risk of a measurable function $f : X \to \mathbb{R}$ by

$$\begin{aligned}\mathcal{R}_{L,\mathrm{P}}(f) &:= \int_{X\times Y} L\big(y, f(x)\big)\, d\mathrm{P}(x, y) \\ &= \int_X \eta(x) L\big(1, f(x)\big) + \big(1 - \eta(x)\big) L\big(-1, f(x)\big)\, d\mathrm{P}_X(x),\end{aligned}$$

where P_X denotes the marginal distribution of P, and $\eta(x) := \mathrm{P}(y = 1|x)$, $x \in X$, the posterior probability of P. Note that these definitions yield $\mathcal{R}_{L,\mathrm{P}}(f) = \mathbb{E}_\mathrm{P} L \circ f$, and depending on the situation we will use either of these notations. Finally, if P is the empirical measure of a sample set $D = ((x_1, y_1), \ldots, (x_n, y_n)) \in (X \times Y)^n$ of length n, we usually write $\mathcal{R}_{L,\mathrm{D}}(f) := \mathcal{R}_{L,\mathrm{P}}(f)$. Analogously, we denote the empirical expectation with respect to D by $\mathbb{E}_\mathrm{D}$.

Throughout this work the smallest possible L-risk

$$\mathcal{R}^*_{L,\mathrm{P}} := \inf\big\{\mathcal{R}_{L,\mathrm{P}}(f) \,\big|\, f : X \to \mathbb{R} \text{ measurable}\big\}$$

is called the Bayes risk with respect to P and L. Furthermore, a measurable function $f^*_{L,\mathrm{P}} : X \to \mathbb{R}$ with $\mathcal{R}_{L,\mathrm{P}}(f^*_{L,\mathrm{P}}) = \mathcal{R}^*_{L,\mathrm{P}}$ is called a Bayes decision function.[1] For example, it is well-known that $f^*_{L_{\text{class}},\mathrm{P}}(x) = \operatorname{sign}(2\eta(x) - 1)$, $x \in X$, is the Bayes decision function for the classification loss. We usually call $f^*_{L_{\text{class}},\mathrm{P}}$ the Bayes classifier.

In the following, we call a Banach space E that consists of functions $f : X \to \mathbb{R}$ a Banach function space (BFS) over X, and we always denote the closed unit ball of E by B_E. Clearly, reproducing kernel Hilbert spaces (RKHSs) are Banach function spaces. In order to introduce another type of BFSs we need the notation

$$\|(w_i)_{i\in I}\|_{\ell_1(I)} := \sum_{i\in I} |w_i|\,,$$

where I is an at most countable and non-empty set, and $(w_i)_{i\in I} \subset \mathbb{R}$ is an $\mathbb{R}$-valued family over I. Clearly, the space

$$\ell_1(I) := \big\{(w_i)_{i\in I} : \|(w_i)_{i\in I}\|_{\ell_1(I)} < \infty\big\}$$

is a separable Banach space. With the help of this space, the following lemma, whose proof can be found in Section 3, introduces the type of BFSs we are most interested in.

Lemma 2.1 *Let I be an at most countable and non-empty set, and $(e_i)_{i\in I}$ be a family of bounded functions $e_i : X \to \mathbb{R}$ with $\|e_i\|_\infty \le 1$ for all $i \in I$. We define*

$$E := \left\{ f : X \to \mathbb{R} \,\middle|\, \exists (w_i)_{i\in I} \in \ell_1(I) \text{ with } f(x) = \sum_{i\in I} w_i e_i(x) \text{ for all } \forall x \in X \right\},$$

[1] Note, that unlike some other authors we demand that Bayes decision functions are *real-valued*, rather than extended real-valued. However, in Subsection 2.2, we will also briefly deal with extended real-valued minimizers.

where we note that the uniform boundedness of the family $(e_i)_{i\in I}$ ensures that the sum above converges absolutely for every $x \in X$. Furthermore, for $f \in E$, we write

$$\|f\|_E := \inf\left\{\sum_{i\in I} |w_i| : (w_i)_{i\in I} \in \ell_1(I) \text{ with } f(x) = \sum_{i\in I} w_i e_i(x) \text{ for all } \forall x \in X\right\}.$$

Then $(E, \|\cdot\|_E)$ is a separable Banach function space that consists of bounded functions and we have

$$\|f\|_\infty \leq \|f\|_E\,, \qquad\qquad f \in E.$$

Finally, if all e_i, $i \in I$, are measurable, then E consists of measurable functions.

Bounds on the generalization performance of regularized empirical risk minimizers often include a complexity measure of the underlying function class. Since in this work we will use average empirical entropy numbers as a complexity measure, let us briefly recall the definition of entropy numbers. To this end, let (T, d) be a metric space and $n \geq 1$ be an integer. Then the n-th (dyadic) entropy number of (T, d) is defined by

$$e_n(T,d) \;:=\; \inf\left\{\varepsilon > 0 : \exists\, s_1, \ldots, s_{2^{n-1}} \in T \text{ such that } T \subset \bigcup_{i=1}^{2^{n-1}} B_d(s_i, \varepsilon)\right\},$$

where we use the convention $\inf \emptyset := \infty$. Moreover, if (T, d) is a subspace of a normed space $(E, \|\cdot\|)$ and the metric d is given by $d(x, x') = \|x - x'\|$, $x, x' \in T$, we write

$$e_n(T, \|\cdot\|) := e_n(T, E) := e_n(T, d)\,.$$

Finally, if $S : E \to F$ is a bounded, linear operator between the normed spaces E and F, we write $e_n(S) := e_n(SB_E, \|\cdot\|_F)$. Entropy numbers are closely related to the well-known covering numbers; in fact, both concepts are inverse to each other modulo constants. We refer to, e.g., [14, Lemma 6.21 & Exercise 6.8] for precise statements and to [6] and [14, Appendix A.5.6] for several properties of entropy numbers. In the following, we are only interested in entropy numbers that are computed with respect to the norm of an empirical L_2-space. To be more precise, let Z be a non-empty set and $D \in Z^n$ be a finite Z-valued sequence of length $n \geq 1$. For $Z \to \mathbb{R}$, we then define

$$\|f\|^2_{L_2(\mathrm{D})} := \frac{1}{n}\sum_{i=1}^{n} |f(z_i)|^2 = \mathbb{E}_{\mathrm{D}}|f|^2\,,$$

and denote the corresponding Hilbert space of equivalence classes by $L_2(\mathrm{D})$. Note that if E is the BFS introduced in Lemma 2.1 and $D_X \in X^n$, then $e_i(\mathrm{id} : E \to L_2(\mathrm{D}_X))$ equals the i-th entropy number of the absolute convex hull of the family $(e_i)_{i\in I}$. By results from, e.g., [7, 5, 9, 12] the latter can be estimated from the entropy numbers of the family $(e_i)_{i\in I}$.

2.2 An Oracle Inequality for Regularized Boosting Algorithms

In this subsection, we establish our first oracle inequality for regularized boosting algorithms and discuss some of its consequences.

Let us begin by formally introducing these learning methods.

Definition 2.2 *Let E be a Banach function space over X and $L : Y \times \mathbb{R} \to [0, \infty)$ be a convex loss function. Given a $\lambda > 0$ and an $\epsilon \geq 0$, we call a learning method that assigns to every $D \in (X \times Y)^n$ a function $f_{D,\lambda} : X \to \mathbb{R}$ such*

$$\lambda \|f_{D,\lambda}\|_E + \mathcal{R}_{L,\mathrm{D}}(f_{D,\lambda}) < \inf_{f \in E} \lambda \|f\|_E + \mathcal{R}_{L,\mathrm{D}}(f) + \epsilon \tag{2}$$

an ϵ-approximate regularized boosting algorithm (ϵ-ARBA) with respect to E and L.

Let us briefly check that the definition above matches our notion of regularized boosting algorithms from the introduction. To this end, we fix the BFS E introduced in Lemma 2.1. For an $w := (w_i)_{i \in I} \in \ell_1(I)$, we further define $f_w := \sum_{i \in I} w_i e_i$. By the definition of $\|\cdot\|_E$ we then immediately obtain

$$\lambda \|f_w\|_E + \mathcal{R}_{L,\mathrm{D}}(f_w) \leq \lambda \sum_{i \in I} |w_i| + \mathcal{R}_{L,\mathrm{D}}(f_w)$$

for all $w \in \ell_1(I)$. Conversely, given an $f \in E$ and an $\varepsilon > 0$, there exists an $w \in \ell_1(I)$ with $f = f_w$ and $\|w\|_{\ell_1(I)} \leq \|f\|_E + \varepsilon$, and hence we find

$$\lambda \sum_{i \in I} |w_i| + \mathcal{R}_{L,\mathrm{D}}(f) \leq \lambda \|f_w\|_E + \mathcal{R}_{L,\mathrm{D}}(f) + \varepsilon \,.$$

From these two inequalities it is straightforward to check that (2) is equivalent to (1). However, our definition of ϵ-ARBAs is not restricted to the BFS of Lemma 2.1. Indeed, if E is a separable RKHS, we obtain a support vector machine (SVM) whose regularization term is not squared. Recall that such SVMs have been recently investigated in [3, 13].

If the BFS E considered in (2) is separable and consists of bounded measurable functions, it is easy to show by an almost literal repetition of the proof of [14, Lemma 6.23] that there exists a measurable version in the sense of [14, Definition 6.2] that satisfies (2). In the following, we *always* implicitly assume that we consider such a measurable version.

We also need infinite sample versions of ϵ-ARBAs. To introduce these, we fix a distribution P on $X \times Y$ and assume that the BFS E over X consists of bounded measurable functions. Then every $f_{\mathrm{P},\lambda} \in E$ satisfying

$$\lambda \|f_{\mathrm{P},\lambda}\|_E + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) < \inf_{f \in E} \lambda \|f\|_E + \mathcal{R}_{L,\mathrm{P}}(f) + \epsilon \tag{3}$$

is called an infinite sample version of the ϵ-ARBA with respect to E and L. Finally, we define the corresponding approximation error function $A : [0, \infty) \to [0, \infty)$ by

$$A(\lambda) := \inf_{f \in E} \lambda \|f\|_E + \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}} \,, \qquad \lambda \geq 0.$$

The following lemma collects some useful properties of the approximation error function. Here we note that the implication from (5) to $A(\lambda) \leq c\lambda$ for all $\lambda \geq 0$ was already observed in [4].

Lemma 2.3 *$L : Y \times \mathbb{R} \to [0,\infty)$ be a convex loss, E be a BFS over X that consists of bounded measurable functions, and P be a distribution on $X \times Y$. Assume that E is sufficiently rich in the sense of $\inf_{f\in E} \mathcal{R}_{L,\mathrm{P}}(f) = \mathcal{R}^*_{L,\mathrm{P}}$. Then the approximation error function $A : [0,\infty) \to [0,\infty)$ is increasing, concave, and continuous. Moreover, we have $A(0) = 0$ and*

$$\begin{aligned} \frac{A(\kappa)}{\kappa} &\leq \frac{A(\lambda)}{\lambda}\,, & 0 < \lambda \leq \kappa, \\ A(\lambda) &\leq \mathcal{R}_{L,\mathrm{P}}(0) - \mathcal{R}^*_{L,\mathrm{P}}\,, & \lambda \geq 0. \end{aligned} \tag{4}$$

In addition, $A(\,\cdot\,)$ is subadditive in the sense of

$$A(\lambda+\kappa) \;\leq\; A(\lambda) + A(\kappa)\,, \qquad \lambda,\kappa \geq 0.$$

Moreover, for a constant $c \geq 0$, we have $A(\lambda) \leq c\lambda$ for all $\lambda \geq 0$ if and only if we have

$$\inf_{f\in cB_E} \mathcal{R}_{L,\mathrm{P}}(f) = \mathcal{R}^*_{L,\mathrm{P}}\,. \tag{5}$$

Finally, if there exists an $h : [0,1] \to [0,\infty)$ with $\lim_{\lambda\to 0^+} h(\lambda) = 0$ and $A(\lambda) \leq \lambda h(\lambda)$ for all $\lambda \in [0,1]$, then we have $A(\lambda) = 0$ for all $\lambda \geq 0$, and $\mathcal{R}_{L,\mathrm{P}}(0) = \mathcal{R}^*_{L,\mathrm{P}}$.

Before we can present our first oracle inequality, we finally need to assume a variance bound. To formulate the latter, we fix a convex, margin-based loss L with $L \neq 0$ and a distribution P on $X \times Y$. We define $\varphi(-\infty) := \lim_{t\to-\infty} \varphi(t)$ and $\varphi(\infty) := \lim_{t\to\infty} \varphi(t)$, where φ is the representing function of L, and extend L to $Y \times [-\infty,\infty]$ in the same way. By the convexity of L it is then easy to show that there exists a measurable function $f^*_{L,\mathrm{P}} : X \to [-\infty,\infty]$ such that $\mathcal{R}_{L,\mathrm{P}}(f^*_{L,\mathrm{P}}) = \mathcal{R}^*_{L,\mathrm{P}}$. Moreover, we can choose $f^*_{L,\mathrm{P}}$ such that $f^*_{L,\mathrm{P}}(x) = \pm\infty$ if and only if $\mathrm{P}(y = 1|x) \in \{0,1\}$. Let us fix such an $f^*_{L,\mathrm{P}}$. Then, we say that L satisfies a variance bound for P, if there exists a constant $c \geq 1$ such that

$$\mathbb{E}_{\mathrm{P}}\big(L\circ f - L\circ f^*_{L,\mathrm{P}}\big)^2 \leq c\,\|L\circ f\|_\infty \big(\mathbb{E}_{\mathrm{P}} L\circ f - L\circ f^*_{L,\mathrm{P}}\big) \tag{6}$$

for all bounded measurable functions $f : X \to \mathbb{R}$. We refer to [1] and [4] for various examples of margin-based losses, including the logistic loss for classification and the AdaBoost loss, that satisfy (6). In particular, recall [4, Lemma 19], which provides an easy sufficient condition for (6) to hold.

With these preparation we can now present our first main result that establishes an oracle inequality for approximate regularized boosting algorithms.

Theorem 2.4 *Let E be a separable Banach function space over X that consists of bounded measurable functions and whose norm satisfies $\|\cdot\|_\infty \leq \|\cdot\|_E$. Moreover, let P be a distribution on $X \times Y$ and L be a margin-based loss that is convex and Lipschitz continuous with $|L|_1 \leq 1$. In addition, assume that its representing function φ satisfies $\varphi(0) \leq 1$. We further assume that P and L satisfy the variance bound (6). We fix an $n \geq 1$ and further assume that there exist constants $a \geq 1$ and $p \in (0,1)$ such that*

$$\mathbb{E}_{D_X\sim \mathrm{P}_X^n} e_i\big(\mathrm{id} : E \to L_2(\mathrm{D}_X)\big) \leq a i^{-\frac{1}{2p}}$$

for all $i \geq 1$. Then there exists a constant $K \geq 1$ only depending on p and c such that, for all $\lambda \in (0,1]$ and $\tau \geq 1$ satisfying $\lambda n \geq K\tau$ and $\lambda^{1+p} n \geq a^{2p} K$, every $\lambda/2$-ARBA with respect to E and L satisfies

$$\mathrm{P}^n\big(D \in (X \times Y)^n : \lambda \|f_{D,\lambda}\|_E + \mathcal{R}_{L,\mathrm{P}}(f_{D,\lambda}) - \mathcal{R}^*_{L,\mathrm{P}} < 2A(\lambda) + \lambda\Big) \geq 1 - e^{-\tau}\,.$$

Note that it is possible to derive a formula for the constant K from the proof of Theorem 2.4. However, the formula has a relatively complicated structure, and in addition, we conjecture, that the resulting values for K are overly pessimistic. Consequently, we omit the details for the sake of simplicity.

One simple way to ensure the average empirical entropy number assumption of Theorem (2.4) is to assume a uniform empirical entropy number assumption. Recall that the latter type of assumption has been widely used in the literature. For example, for RKHSs, the smoothness of the corresponding kernel can ensure such an entropy bound, see, e.g., [14, Theorem 6.26]. Moreover, for the BFSs considered in Lemma 2.1, [1, 4] bound these entropy numbers in terms of the VC-dimension of the family $(e_i)_{i \in I}$. Finally note that although these approaches are easy to use, they may, however, be sometimes not tight. We refer to [14, Theorem 7.34] for an example in this direction.

Let us now briefly illustrate the consequences of the above oracle inequality for ε-ARBAs that uses the BFS of Lemma 2.1. To this end, we fix a sequence $(\lambda_n) \subset (0,1]$ such that $\lambda_n^{1+p} n \geq a^{2p} K$ for all sufficiently large $n \geq 1$. For example, we can choose $\lambda_n := n^{-\frac{1}{1+p}} \ln(n+1)$ if we do not have good estimates for the values of a and K. If the BFS E is rich in the sense of $\inf_{f \in E} \mathcal{R}_{L,\mathrm{P}}(f) = \mathcal{R}^*_{L,\mathrm{P}}$ for all distributions P on $X \times Y$, then the λ_n-ARBA is universally consistent with respect to the risk $\mathcal{R}_{L,\mathrm{P}}$. Moreover, if L is classification calibrated in the sense of [1], i.e. $\varphi'(0) < 0$, then the λ_n-ARBA is also universally classification consistent. Finally, note that [14, Theorem 5.31], see also [4, Lemma 16], shows that the richness assumption above is satisfied if E is dense in $L_1(\mu)$ for all distributions μ on X, and by [14, Theorem 5.36] one can show that for many loss functions the converse implication is also true. In particular, the logistic loss for classification is such a loss.

Let us now assume that there exists constants $c > 0$ and $\beta \in (0,1]$ such that $A(\lambda) \leq c\lambda^\beta$ for all $\lambda > 0$. The sequence (λ_n) considered above then yields the learning rate $n^{-\frac{\beta}{1+p}}$ for the $\mathcal{R}_{L,\mathrm{P}}$-risks of the ARBA. Recall that [1] showed that this leads to the learning rate $n^{-\frac{\beta}{2+2p}}$ for the classification risk, and if the distribution satisfies Tsybakov's noise assumption, see [16], this rate can be improved up to the rate $n^{-\frac{\beta}{1+p}}$. We refer to [1] and [14, Chapter 3 & Chapter 8] for details. In any case, these learning rates coincide with the learning rates established in [4] for certain discretized versions of ARBAs, where we refer to [4, page 884] for the necessary translation of VC-dimension bounds to entropy number bounds. Finally, [4, Corollary 9] shows that these learning rates are asymptotically optimal if E is built from decision stumps, the logistic loss is used, and the function $f^*_{L,\mathrm{P}} : X \to [-\infty, \infty]$ is of bounded variation. Note that the latter implies that the posterior probability $\eta : X \to [0,1]$ is bounded away from zero and one and that $\beta = 1$. In order to artificially enforce the former, [4] suggests to add a coin flipping noise to the labels. In the following subsection, we will establish an oracle inequality for this approach, which, for $\beta < 1$, leads to improved

learning rates.

2.3 Oracle Inequalities for two-sided losses

As mentioned at the end of the previous subsection, our goal in this subsection is to establish an oracle inequality that addresses the idea of adding a coin flipping noise to the labeling process. The key technique to establish this oracle inequality is to translate this additive noise into a loss function that enjoys additional properties. With the help of these properties we can then refine our analysis in the case $\beta < 1$.

Let us begin by introducing some more notions. Following [14], we say that a loss L can be clipped at $M > 0$ if, for all $(y, t) \in Y \times \mathbb{R}$, we have

$$L(y, \hat{t}) \;\leq\; L(y, t)\,,$$

where $\hat{t}$ denotes the clipped value of t at $\pm M$, that is

$$\hat{t} := \begin{cases} -M & \text{if } t < -M \\ t & \text{if } t \in [-M, M] \\ M & \text{if } t > M\,. \end{cases} \tag{7}$$

Moreover, we say that L can be clipped if it can be clipped at some $M > 0$. Informally speaking, losses that can be clipped, allow us to restrict our consideration to prediction values between $-M$ and M. With the help of [14, Lemma 2.23] it is easy to check that a margin based loss L can be clipped if and only if its representing function φ has a global minimum. If φ is continuous, we can then choose M to be the smallest value at which this global minimum is attained.

The following lemma gives a simple criterion when a convex, margin-based loss has a Bayes decision function for all distributions P on $X \times Y$.

Lemma 2.5 *Let L be a convex, margin-based loss. Then L can be clipped, if and only if there exists a Bayes decision function $f^*_{L,\mathrm{P}} : X \to \mathbb{R}$ for all distributions P on $X \times Y$. In this case, there always exists a Bayes decision function $f^*_{L,\mathrm{P}} : X \to [-M, M]$, where $M > 0$ is a real number at which L can be clipped.*

Obviously, neither the logistic loss nor AdaBoost loss can be clipped, and it is well-known, that they fail to have a Bayes decision function for exactly the distributions P that have a noise-free region for the labeling process, i.e,

$$\mathrm{P}_X(\{x : \eta(x) = 0 \text{ or } \eta(x) = 1\}) > 0\,. \tag{8}$$

On the other hand, if we have a convex, margin-based loss L it is not hard to see that there exists a Bayes decision function $f^*_{L,\mathrm{P}}$ for all distributions P on $X \times Y$ that do *not* satisfy (8), i.e., for distributions that are noisy everywhere. In addition, this Bayes decision function is P_X-almost surely determined if L is strictly convex. For example, for the logistic loss, we have

$$f^*_{L,\mathrm{P}}(x) = \ln \frac{\eta(x)}{1 - \eta(x)}\,, \qquad\qquad x \in X,$$

and for the AdaBoost loss we have $f^*_{L,\mathrm{P}}(x) = \frac{1}{2}\ln\frac{\eta(x)}{1-\eta(x)}$, $x \in X$. Clearly, if $\eta(x)$ approaches 0 or 1, these Bayes decision functions become unbounded as we have already used at the end of Subsection 2.2. Since by [14, Corollary 3.62] every reasonable learning algorithm based on these loss functions has to approximate $f^*_{L,\mathrm{P}}$, we see that such algorithms have to approximate potentially unbounded functions. In order to avoid such a behavior, a commonly used trick, see e.g., [17, page 2280] and [4, page 873], is to add some noise to the labeling process. More precisely, if P is a distribution on $X \times Y$ with marginal distribution P_X and posterior probability η, and $0 < \delta < 1/2$, we can define a new distribution $\mathrm{P}^{(\delta)}$ on $X \times Y$ by $\mathrm{P}^{(\delta)}_X := \mathrm{P}_X$ and

$$\eta^{(\delta)}(x) := (1-\delta)\eta(x) + \delta\big(1-\eta(x)\big)\,, \qquad\qquad x \in X.$$

In other words, $\mathrm{P}^{(\delta)}$ is constructed by adding some noise of order δ to the posterior probability η of P. Now note that we have $1/2 < \eta^{(\delta)}(x) < \eta(x)$, if $\eta(x) > 1/2$, and $\eta(x) < \eta^{(\delta)}(x) < 1/2$, if $\eta(x) < 1/2$. Consequently, the Bayes classifiers of both distributions P and $\mathrm{P}^{(\delta)}$ coincide. Moreover, it is easy to see that $\delta \le \eta^{(\delta)}(x) \le 1-\delta$, i.e., $\mathrm{P}^{(\delta)}$ is noisy everywhere. In particular, all convex and margin-based losses have a Bayes decision function for $\mathrm{P}^{(\delta)}$.

Our next goal is to encode the above construction into a loss function. To this end, we need the following definition.

Definition 2.6 *Let L be a loss function and $0 < \delta < 1/2$. Then the δ-two-sided version $L_\delta : Y \times \mathbb{R} \to [0,\infty)$ of L is defined by*

$$L_\delta(y,t) := (1-\delta)L(y,t) + \delta L(-y,t)\,, \qquad\qquad y \in Y,\ t \in \mathbb{R}.$$

Note that, for every distribution P on $X \times Y$ and every measurable $f : X \to \mathbb{R}$, a straightforward calculation, see (16) and (17), shows

$$\mathcal{R}_{L_\delta,\mathrm{P}}(f) = \mathcal{R}_{L,\mathrm{P}^{(\delta)}}(f)\,. \tag{9}$$

In other words, adding some noise to the posterior probabilities is, in terms of the learning goals described by the risk functionals, equivalent to using the two-sided version of a loss function. However, in terms of algorithmic design, there may be a substantial difference in both approaches. Indeed, a straightforward implementation of using $\mathrm{P}^{(\delta)}$ would individually flip each label y_i of the training set with probability δ whereas an algorithm based on L_δ pretends to see y_i with probability $(1-\delta)$ and $-y_i$ with probability δ, simultaneously.

Before we present our oracle inequality for algorithms based on L_δ we need a few more preparations. Let us begin with the following lemma that collects some simple, yet useful properties of two-sided versions of margin based losses.

Lemma 2.7 *Let L be a convex, margin-based loss and δ a real number with $0 < \delta < 1/2$. Then the following statements are true for the δ-two-sided version L_δ of L:*

i) L_δ is Lipschitz continuous, or strictly convex if and only if L is.

ii) L_δ *can be clipped at*

$$M_\delta := \inf\bigl\{t \in \mathbb{R} : L_\delta(1,t) \le L_\delta(1,s) \text{ for all } s \in \mathbb{R}\bigr\}. \tag{10}$$

iii) For every probability measure P *on* $X \times Y$ *there exists a Bayes decision function* $f^*_{L_\delta,\mathrm{P}} : X \to [-M_\delta, M_\delta]$. *In addition, if* L *is strictly convex,* $f^*_{L_\delta,\mathrm{P}}$ *is uniquely determined and we have*

$$f^*_{L_\delta,\mathrm{P}} = f^*_{L,\mathrm{P}^{(\delta)}}\,.$$

With the help of the lemma above we can now introduce the learning methods we consider in this subsection.

Definition 2.8 *Let* E *be a Banach function space over* X, $L : Y \times \mathbb{R} \to [0,\infty)$ *be a convex, margin-based loss function, and* L_δ *its* δ*-two-sided version for some* $0 < \delta < 1/2$. *Given a* $\lambda > 0$ *and an* $\epsilon \ge 0$, *we call a learning method that assigns to every* $D \in (X \times Y)^n$ *a function* $f_{D,\lambda} : X \to \mathbb{R}$ *such*

$$\lambda\|f_{D,\lambda}\|_E + \mathcal{R}_{L,\mathrm{D}}(\widehat{f_{D,\lambda}}) < \inf_{f\in E} \lambda\|f\|_E + \mathcal{R}_{L,\mathrm{D}}(f) + \epsilon \tag{11}$$

a clipped ϵ*-approximate regularized boosting algorithm (*ϵ*-CARBA) with respect to* E *and* L. *Here, the clipping operation* $\widehat{\cdot}$ *is with respect to* M_δ *defined in (10).*

CARBAs are a particular example of more general, clipped regularized empirical risk minimizers introduced in [14, Chapter 7.4]. We refer to this chapter for a discussion of these learning methods including the existence of measurable versions.

Before we can establish an oracle inequality for CARBAs we finally need to present a variance bound for two-sided losses. This is done in the following proposition, which extends [4, Lemma 19].

Proposition 2.9 *Let* L *be a strictly convex, twice continuously differentiable, classification calibrated, and margin-based loss. We fix a* δ *with* $0 < \delta < 1/2$ *and define* M_δ *by (10). We further write*

$$\tilde{C}_L(\delta) := \sup\left\{ \frac{2\varphi'(t)\varphi'(-t)(\varphi'(t)+\varphi'(-t))}{\varphi'(-t)\varphi''(t)+\varphi'(t)\varphi''(-t)} - \varphi(t) - \varphi(-t) : t \in [-M_\delta, M_\delta] \right\}$$

and $C_L(\delta) := \max\{0, \tilde{C}_L(\delta)\}$. *Then for all distributions* P *on* $X \times Y$ *and all measurable functions* $f : X \to [-M_\delta, M_\delta]$ *we have*

$$\mathbb{E}_\mathrm{P}\bigl(L_\delta \circ f - L \circ f^*_{L_\delta,\mathrm{P}}\bigr)^2 \le \bigl(\varphi(M_\delta) + \varphi(-M_\delta) + C_L(\delta)\bigr)\mathbb{E}_\mathrm{P}\bigl(L_\delta \circ f - L \circ f^*_{L_\delta,\mathrm{P}}\bigr).$$

Note that the strict convexity and differentiability of φ is actually only needed on the interval $[-M_\delta, M_\delta]$, if the Bayes decision function $f^*_{L_\delta,\mathrm{P}}$ is uniquely determined. Moreover, the same is true for the following theorem, which presents the already announced oracle inequality for CARBAs.

Theorem 2.10 *Let E be a separable Banach function space over X that consists of bounded measurable functions and whose norm satisfies $\|\cdot\|_\infty \leq \|\cdot\|_E$. Moreover, let P be a distribution on $X \times Y$ and L be a margin-based loss that is strictly convex, twice continuously differentiable, classification calibrated and Lipschitz continuous with $|L|_1 \leq 1$. In addition, assume that its representing function φ satisfies $\varphi(0) \leq 1$. We fix an $n \geq 1$ and further assume that there exist constants $a \geq \varphi(-M_\delta)$ and $p \in (0,1)$ such that*

$$\mathbb{E}_{D_X \sim \mathrm{P}_X^n} e_i\bigl(\mathrm{id} : E \to L_2(\mathrm{D}_X)\bigr) \leq a i^{-\frac{1}{2p}}$$

for all $i \geq 1$. Then there exists a constant $K \geq 1$ only depending on L, δ, and p such that, for all $\lambda \in (0,1]$, $\epsilon \geq 0$, and $\tau > 0$, every ϵ-CARBA with respect to E and L satisfies with probability P^n not less than $1 - e^{-\tau}$

$$\lambda \|\widehat{f}_{D,\lambda}\|_E + \mathcal{R}_{L_\delta,\mathrm{P}}(f_{D,\lambda}) - \mathcal{R}^*_{L_\delta,\mathrm{P}} < 15A(\lambda) + K\Bigl(\frac{a^{2p}}{\lambda^{2p} n}\Bigr)^{\frac{1}{1-p}} + K\frac{\tau}{n} + \frac{30A(\lambda)}{\lambda n} + 3\epsilon\,,$$

where the approximation error function is with respect L_δ.

Let us now briefly compare the consequences of Theorem 2.10 with those of Theorem 2.4. To this end, we again consider a BFS E defined by Lemma 2.1. Obviously, Theorem 2.10 yields consistency of λ_n-CARBAs if we fix a sequence $(\lambda_n) \subset (0,1]$ with $\lambda_n \to 0$, $\lambda_n^{2p} n \to \infty$, and $\lambda_n n \geq 1$ for all $n \geq 1$. Moreover, if we assume that there exist constants $c \geq 0$ and $\beta \in (0,1]$ such that $A(\lambda) \leq c\lambda^\beta$ for all $\lambda > 0$, then [14, Lemma A.1.7] shows that $\lambda_n := n^{-\rho}$, where[2]

$$\rho := \min\Bigl\{1, \frac{1}{\beta(1-p)+2p}\Bigr\}\,,$$

asymptotically minimizes the right hand side of the oracle inequality of Theorem 2.10. Obviously, this yields the learning rate $n^{-\rho\beta}$ with respect to the risk $\mathcal{R}_{L_\delta,\mathrm{P}}$, and by (9), this rate can be immediately translated into a learning rate for binary classification. Analogously to the learning rates derived from Theorem 2.4, these learning rates can be further improved if P, or equivalently $\mathrm{P}^{(\delta)}$, satisfies a Tsybakov noise assumption.

Finally note that *formally* the learning rates $n^{-\rho\beta}$ are faster than those derived from Theorem 2.4, whenever $\beta < 1$, while for $\beta = 1$ both learning rates coincide. However, strictly speaking we cannot compare both learning rates since they are based on assumptions on *different* approximation error functions. In this direction we note that the decision stumps considered by [4] only yield $\beta = 1$ for the logistic loss L if $x \mapsto \ln \frac{\eta(x)}{1-\eta(x)}$ has bounded variation. In particular, $\eta(x)$ must be bounded away from both 0 and 1. On the other hand, it is easy to check that the approximation error function for a two-sided version L_δ of L satisfies $A(\lambda) \leq c\lambda$ for a constant $c \geq 0$ and all $\lambda > 0$, if $x \mapsto \eta(x)$ has bounded variation. In particular, it is *not* necessary that $\eta(x)$ is bounded away from zero and one, i.e., the faster learning rate of Theorem 2.10 holds under weaker assumptions on P. We conjecture, that this relationship between the two approximation error functions holds in most situations.

[2]Note that this choice of λ_n obviously requires knowledge on β, which, in general, is not available. However, since L_δ can be clipped, the adaptive training/validation approach of [14, Chapter 7.4] can be easily modified to CARBAs.

3 Proofs

Proof of Lemma 2.1: We obviously have $\|f\|_E \geq 0$ for all $f \in E$, and it is also obvious that $\|f\|_E = 0$ if and only if $f = 0$. Now let $f, g \in E$ have the representations $f = \sum_{i\in I} w_i e_i$ and $g = \sum_{i\in I} v_i e_i$, where we note that our assumptions guarantee that the sums converge pointwise absolutely. We then find $f + g = \sum_{i\in I}(w_i + v_i)e_i$, and hence we conclude

$$\|f + g\|_E \leq \sum_{i\in I} |w_i + v_i| \leq \sum_{i\in I} |w_i| + \sum_{i\in I} |v_i| \,.$$

Considering the infimum over all representations of f and g, we then obtain the triangle inequality for $\|\cdot\|_E$. The homogeneity of $\|\cdot\|_E$, i.e., $\|\alpha f\|_E = |\alpha|\cdot\|f\|_E$ can be shown analogously.

Let us now show that $\|\cdot\|_E$ is complete[3]. To this end, we fix sequence $(f_j)_{j\geq 1} \subset E$ with $\sum_{j=1}^{\infty} \|f_j\|_E < \infty$. Moreover, for all $j \geq 1$, we fix a representation

$$f_j = \sum_{i\in I} w_i^{(j)} e_i$$

such that $\sum_{i\in I} |w_i^{(j)}| \leq \|f_j\|_E + 2^{-j}$. For $i_0 \in I$ we then have

$$\sum_{j=1}^{\infty} |w_{i_0}^{(j)}| \leq \sum_{j=1}^{\infty} \sum_{i\in I} |w_i^{(j)}| \leq \sum_{j=1}^{\infty} \bigl(\|f_j\|_E + 2^{-j}\bigr) < \infty \,, \tag{12}$$

and hence $w_{i_0} := \sum_{j=1}^{\infty} w_{i_0}^{(j)}$ does exists. Moreover, by ignoring the first inequality on the left hand side of (12), we further see that (12) yields $(w_i)_{i\in I} \in \ell_1(I)$. Let us now define $f := \sum_{i\in I} w_i e_i \in E$, where we note that this sums also converges pointwise absolutely since $(w_i)_{i\in I} \in \ell_1(I)$ and all e_i are assumed to be bounded functions. Consequently, for all $x \in X$, we have

$$f(x) - \sum_{j=1}^{n} f_j(x) = \sum_{i\in I} \Bigl(w_i - \sum_{j=1}^{n} w_i^{(j)}\Bigr) e_i(x) = \sum_{i\in I} \sum_{j=n+1}^{\infty} w_i^{(j)} e_i(x) \,,$$

and from this we deduce

$$\Bigl\| f - \sum_{j=1}^{n} f_j \Bigr\|_E \leq \sum_{i\in I} \Bigl| \sum_{j=n+1}^{\infty} w_i^{(j)} \Bigr| \leq \sum_{j=n+1}^{\infty} \sum_{i\in I} |w_i^{(j)}| \leq \sum_{j=n+1}^{\infty} \bigl(\|f_j\|_E + 2^{-j}\bigr) \leq \varepsilon$$

for all sufficiently large $n \in \mathbb{N}$. In other words, we have found $f = \sum_{j=1}^{\infty} f_j$, where the convergence is with respect to $\|\cdot\|_E$. From this we easily deduce the completeness of $\|\cdot\|_E$.

The separability of E is trivial, and so is the fact that the measurability of all e_i implies the measurability of all $f \in E$.

[3]One could shorten this part of the proof by using the fact the $\Psi : \ell_1(I) \to E$ defined by $(w_i)_{i\in I} \mapsto \sum_{i\in I} w_i e_i$ is, by definition, a metric surjection. However, we preferred to present an elementary proof.

Finally, in order to show that E consists of bounded functions, we fix an $f \in E$ and a representation $f = \sum_{i \in I} w_i e_i$. We then obtain

$$\|f\|_\infty = \Big\| \sum_{i \in I} w_i e_i \Big\|_\infty \leq \sum_{i \in I} |w_i|$$

and by taking the infimum over all representations, we thus find $\|f\|_\infty \leq \|f\|_E$. ■

3.1 Proofs of the results related to Theorem 2.4

Proof of Lemma 2.3: Besides the equivalence related to (5) all assertions follow by a literal repetition of the proof of [14, Lemma 5.15]. Let us now assume that we have $A(\lambda) \leq c\lambda$ for a constant $c > 0$ and all $\lambda > 0$. We fix a $\lambda > 0$ and an $\varepsilon > 0$, and define $\delta := \varepsilon c \lambda$. There then exists an $f_{\lambda,\varepsilon}$ such that

$$\lambda \|f_{\lambda,\varepsilon}\|_E \leq \lambda \|f_{\lambda,\varepsilon}\|_E + \mathcal{R}_{L,\mathrm{P}}(f_{\lambda,\varepsilon}) - \mathcal{R}^*_{L,\mathrm{P}} \leq A(\lambda) + \delta \leq (1+\varepsilon) c \lambda \,,$$

and hence we conclude $f_{\lambda,\varepsilon} \in (1+\varepsilon) c B_E$. For fixed $\varepsilon > 0$ and $\lambda \to 0$ we further conclude $\mathcal{R}_{L,\mathrm{P}}(f_{\lambda,\varepsilon}) \to \mathcal{R}^*_{L,\mathrm{P}}$, and therefore we find

$$\inf_{f \in (1+\varepsilon) c B_E} \mathcal{R}_{L,\mathrm{P}}(f) = \mathcal{R}^*_{L,\mathrm{P}}$$

for all $\varepsilon > 0$. By letting $\varepsilon \to 0$ we then obtain (5). Conversely, if (5) holds, there exists, for all $\varepsilon > 0$, an $f_\varepsilon \in c B_E$ such that $\mathcal{R}_{L,\mathrm{P}}(f_\varepsilon) \leq \mathcal{R}^*_{L,\mathrm{P}} + \varepsilon$. Consequently, we find

$$A(\lambda) \leq \lambda \|f_\varepsilon\|_E + \mathcal{R}_{L,\mathrm{P}}(f_\varepsilon) - \mathcal{R}^*_{L,\mathrm{P}} \leq c\lambda + \varepsilon$$

for all $\lambda > 0$ and $\varepsilon > 0$. By letting $\varepsilon \to 0$, we then obtain $A(\lambda) \leq c\lambda$ for all $\lambda > 0$. ■

In the following lemmas, we consider ϵ-approximate regularized boosting algorithms and their infinite sample counterparts. Before we can formulate these lemmas we need to introduce one more notation. To this end, we fix a separable BFS E over X that consists of bounded measurable functions, a loss function $L : Y \times \mathbb{R} \to [0,\infty)$, a $\lambda > 0$, and an $\epsilon \in [0,1]$. Then for all $f \in E$ we define $g_{f,\lambda} : X \times Y \to \mathbb{R}$ by

$$g_{f,\lambda}(x,y) := \lambda \|f\|_E + L(y, f(x)) - \lambda \|f_{\mathrm{P},\lambda}\| - L(y, f_{\mathrm{P},\lambda}(x)) \,, \tag{13}$$

where $f_{\mathrm{P},\lambda}$ denotes an arbitrary but *fixed* function satisfying (3) for some fixed $\epsilon \in [0,1]$.

The following lemma, which is needed for the proof of Theorem 2.4, establishes a supremum bound on $g_{f,\lambda}$.

Lemma 3.1 *Let E be a separable BFS over X that consists of bounded measurable functions and whose norm satisfies $\|\cdot\|_\infty \leq \|\cdot\|_E$. Moreover, let P be a distribution on $X \times Y$ and L be a convex, Lipschitz continuous, and margin-based loss whose representing function $\varphi : \mathbb{R} \to [0,\infty)$ satisfies $\varphi(0) \leq 1$, and whose Lipschitz constant satisfies $|L|_1 \leq 1$. Then for all $0 < \lambda \leq 1$ and $f \in E$ we have*

$$\|f\|_E \leq \frac{A(\lambda) + \mathbb{E}_{\mathrm{P}} g_{f,\lambda} + \epsilon}{\lambda} \tag{14}$$

$$\|g_{f,\lambda}\|_\infty \leq 4 \cdot \frac{A(\lambda) + \lambda + \mathbb{E}_{\mathrm{P}} g_{f,\lambda} + \epsilon}{\lambda} \tag{15}$$

Proof: Let us fix an $f \in E$. Then we have

$$\begin{aligned}\lambda\|f\|_E &\leq \lambda\|f\|_E + \mathcal{R}_{L,\mathrm{P}}(f) - \mathcal{R}^*_{L,\mathrm{P}} \\ &\leq \lambda\|f_{\mathrm{P},\lambda}\|_E + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) - \mathcal{R}^*_{L,\mathrm{P}} + \mathbb{E}_{\mathrm{P}} g_{f,\lambda} \\ &\leq A(\lambda) + \mathbb{E}_{\mathrm{P}} g_{f,\lambda} + \epsilon\,,\end{aligned}$$

and hence (14) follows. In order to show (15), we first observe that the Lipschitz continuity of φ together with $\varphi(0) \leq 1$ implies $L(y,t) \leq 1 + |t|$ for all $y \in Y$ and $t \in \mathbb{R}$. By $\|\cdot\|_\infty \leq \|\cdot\|_E$ and (14), we consequently obtain

$$\begin{aligned}\big\|\,\lambda\|f\|_E + L\circ f\,\big\|_\infty &\leq \lambda\|f\|_E + 1 + \|f\|_\infty \\ &\leq A(\lambda) + \mathbb{E}_{\mathrm{P}} g_{f,\lambda} + \epsilon + 1 + \frac{A(\lambda) + \mathbb{E}_{\mathrm{P}} g_{f,\lambda} + \epsilon}{\lambda} \\ &\leq 2\cdot\frac{A(\lambda) + \lambda + \mathbb{E}_{\mathrm{P}} g_{f,\lambda} + \epsilon}{\lambda}\,,\end{aligned}$$

where in the last step we used $0 < \lambda \leq 1$. Since this inequality holds for all $f \in E$, we then obtain the assertion. ∎

The following lemma translates the variance bound (6) into a bound on $\mathbb{E}_{\mathrm{P}} g^2_{f,\lambda}$.

Lemma 3.2 *Let E be a separable BFS over X that consists of bounded measurable functions and whose norm satisfies $\|\cdot\|_\infty \leq \|\cdot\|_E$. Moreover, let P be a distribution on $X \times Y$ and L be a convex, Lipschitz continuous, and margin-based loss whose representing function $\varphi : \mathbb{R} \to [0,\infty)$ satisfies $\varphi(0) \leq 1$, and whose Lipschitz constant satisfies $|L|_1 \leq 1$. Assume that the variance bound (6) holds. Then for all $\lambda \in (0,1]$ and all $f \in 2\lambda^{-1}B_E$ we have*

$$\mathbb{E}_{\mathrm{P}} g^2_{f,\lambda} \leq 12\,c\,\lambda^{-1}\big(A(\lambda) + \lambda + \mathbb{E}_{\mathrm{P}} g_{f,\lambda} + \epsilon\big)^2\,.$$

Proof: We fix a $\lambda \in (0,1]$ and an $f \in 2\lambda^{-1}B_E$. Using the shorthands $\mathbb{E} := \mathbb{E}_{\mathrm{P}}$, $g := g_{f,\lambda}$, and $\|\cdot\| = \|\cdot\|_E$, as well as $(a_1+a_2+a_3)^2 \leq 3a_1^2 + 3a_2^2 + 3a_3^2$ for all $a_1, a_2, a_3 \geq 0$, we then obtain

$$\begin{aligned}\mathbb{E}g^2 &= \mathbb{E}\big(\lambda\|f\| - \lambda\|f_{\mathrm{P},\lambda}\| + L\circ f - L\circ f_{\mathrm{P},\lambda}\big)^2 \\ &\leq 3\lambda^2\|f\|^2 + 3\lambda^2\|f_{\mathrm{P},\lambda}\|^2 + 3\mathbb{E}\big(L\circ f - L\circ f_{\mathrm{P},\lambda}\big)^2 \\ &\leq 6\mathbb{E}\big(L\circ f - L\circ f^*_{L,\mathrm{P}}\big)^2 + 6\mathbb{E}\big(L\circ f^*_{L,\mathrm{P}} - L\circ f_{\mathrm{P},\lambda}\big)^2 + 3\lambda^2\|f\|^2 + 3\lambda^2\|f_{\mathrm{P},\lambda}\|^2,\end{aligned}$$

Let us write $C := \max\left(\|f\|_\infty + 1, \|f_{\mathrm{P},\lambda}\|_\infty + 1\right)$. Then the assumption (6) implies

$$\begin{aligned}&\mathbb{E}\big(L\circ f - L\circ f^*_{L,\mathrm{P}}\big)^2 + \mathbb{E}\big(L\circ f^*_{L,\mathrm{P}} - L\circ f_{\mathrm{P},\lambda}\big)^2 \\ \leq\;& c\,C\left(\mathbb{E}\big(L\circ f - L\circ f^*_{L,\mathrm{P}}\big) + \mathbb{E}\big(L\circ f_{\mathrm{P},\lambda} - L\circ f^*_{L,\mathrm{P}}\big)\right).\end{aligned}$$

By assumption we further have $\lambda\|f\| \leq 2$, and since $\varphi(0) \leq 1$ and $\epsilon \in [0,1]$ we also have

$$\lambda\|f_{\mathrm{P},\lambda}\| \leq \lambda\|f_{\mathrm{P},\lambda}\| + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) \leq \mathcal{R}_{L,\mathrm{P}}(0) + \epsilon \leq 1 + \epsilon \leq 2\,.$$

Combining these estimates, we thus obtain

$$\begin{aligned}
\mathbb{E}g^2 &\leq 6\,c\,C\left(\mathbb{E}\big(L\circ f - L\circ f^*_{L,\mathrm{P}}\big)+\mathbb{E}\Big(L\circ f_{\mathrm{P},\lambda} - L\circ f^*_{L,\mathrm{P}}\Big)\right)+3\lambda^2\|f\|^2+3\lambda^2\|f_{\mathrm{P},\lambda}\|^2 \\
&\leq 6\,c\,C\left(\mathbb{E}\big(L\circ f - L\circ f^*_{L,\mathrm{P}}\big)+\mathbb{E}\big(L\circ f_{\mathrm{P},\lambda} - L\circ f^*_{L,\mathrm{P}}\big)+\lambda\|f\|+\lambda\|f_{\mathrm{P},\lambda}\|\right) \\
&= 6\,c\,C\left(\mathbb{E}g + 2\mathbb{E}\big(L\circ f_{\mathrm{P},\lambda} - L\circ f^*_{L,\mathrm{P}}\big) + 2\lambda\|f_{\mathrm{P},\lambda}\|\right) \\
&\leq 12\,c\,C\left(A(\lambda)+\lambda+\mathbb{E}g+\epsilon\right).
\end{aligned}$$

Let us finally bound the constant C. To that end, observe that Lemma 3.1 implies

$$\|f\|_\infty + 1 \leq \|f\|_E + 1 \leq \frac{A(\lambda)+\lambda+\mathbb{E}_{\mathrm{P}}g_{f,\lambda}+\epsilon}{\lambda}$$

for all $f \in E$. Combining this estimate with our previous considerations then yields the assertion. ∎

Finally, we need to translate the entropy number bound assumed in Theorem 2.4 into an entropy number bound on certain sets of functions of the form $g_{f,\lambda}$.

Lemma 3.3 *Let E be a separable Banach function space over X that consists of bounded measurable functions and whose norm satisfies $\|\cdot\|_\infty \leq \|\cdot\|_E$. Moreover, let P be a distribution on $X \times Y$ and L be a Lipschitz continuous and margin-based loss whose representing function $\varphi : \mathbb{R} \to [0,\infty)$ satisfies $\varphi(0) \leq 1$, and whose Lipschitz constant satisfies $|L|_1 \leq 1$. Assume that for a fixed $n \geq 1$ there exist constants $a \geq 1$ and $p \in (0,1)$ such that*

$$\mathbb{E}_{D_X\sim \mathrm{P}_X^n} e_i\big(\mathrm{id} : E \to L_2(\mathrm{D}_X)\big) \leq a i^{-\frac{1}{2p}}\,, \qquad\qquad i \geq 1.$$

For $\lambda \in (0,1]$ and $\varepsilon > 0$ with $\varepsilon \leq A(\lambda)+\lambda+\epsilon$, we define

$$\mathcal{G}_\varepsilon(\lambda) := \left\{g_{f,\lambda} : f \in 2\lambda^{-1}B_E \text{ and } \mathbb{E}_{\mathrm{P}}g_{f,\lambda} \leq \varepsilon\right\}.$$

Then we have

$$\mathbb{E}_{D\sim \mathrm{P}^n} e_i\big(\mathcal{G}_\varepsilon(\lambda), \|\cdot\|_{L_2(\mathrm{D})}\big) \leq c_p\, a\, \frac{A(\lambda)+\lambda+\varepsilon+\epsilon}{\lambda}\, i^{-\frac{1}{2p}}\,, \qquad\qquad i \geq 1.$$

Proof: Let us fix a $\lambda \in (0,1]$ and an $\varepsilon > 0$. For $\Lambda := \frac{A(\lambda)+\lambda+\varepsilon+\epsilon}{\lambda}$ we then observe by Lemma 3.1 that $\|f\|_E \leq \Lambda$ for all $f \in \mathcal{G}_\varepsilon(\lambda)$. Moreover, $\lambda \leq 1$, $\epsilon \leq 1$, $A(\lambda) \leq 1$, see Lemma 2.3, and $\varepsilon \leq A(\lambda)+\lambda+\epsilon$, implies $\Lambda \leq 6\lambda^{-1}$. Let us now define the auxiliary sets

$$\begin{aligned}
\mathcal{G} &:= \{\lambda\|f\|_E + L\circ f : f \in \Lambda B_E\}, \\
\mathcal{R} &:= \{\lambda\|f\|_E : f \in \Lambda B_E\}, \\
\mathcal{H} &:= \{L\circ f : f \in \Lambda B_E\}.
\end{aligned}$$

The translation invariance and additivity of the entropy numbers, see the arguments on pages 11 & 12 of [6] for the latter, then yields

$$\begin{aligned}
\mathbb{E}_{D\sim \mathrm{P}^n} e_{2i-1}\bigl(\mathcal{G}_\varepsilon(\lambda), \|\cdot\|_{L_2(\mathrm{D})}\bigr) &\leq \mathbb{E}_{D\sim \mathrm{P}^n} e_{2i-1}\bigl(\mathcal{G}, \|\cdot\|_{L_2(\mathrm{D})}\bigr) \\
&\leq \mathbb{E}_{D\sim \mathrm{P}^n}\Bigl(e_i\bigl(\mathcal{R}, \|\cdot\|_{L_2(\mathrm{D})}\bigr) + e_i\bigl(\mathcal{H}, \|\cdot\|_{L_2(\mathrm{D})}\bigr)\Bigr) \\
&\leq e_i\bigl([0,6], |\cdot|\bigr) + \Lambda \mathbb{E}_{D_X\sim \mathrm{P}_X^n} e_i\bigl(B_E, \|\cdot\|_{L_2(\mathrm{D}_X)}\bigr) \\
&\leq 3\cdot 2^{-i} + a\Lambda i^{-\frac{1}{2p}} \\
&\leq \tilde{c}_p a \Lambda (2i-1)^{-\frac{1}{2p}}\,,
\end{aligned}$$

where $\tilde{c}_p$ is a constant only depending on p. By the monotonicity of the entropy numbers, we then also find the assertion for even indices, if we increase the constant $\tilde{c}_p$ by a factor only depending on p. ∎

Proof of Theorem 2.4: We write $\mathcal{G}(\lambda) := \{g_{f,\lambda} : f \in 2\lambda^{-1}B_E\}$ and define $\mathcal{G}_\varepsilon(\lambda)$ as in Lemma 3.3. Moreover, we write

$$\Lambda(\lambda,\varepsilon) := \frac{A(\lambda)+\lambda+\varepsilon+\epsilon}{\lambda}$$

for all $\lambda \in (0,1]$ and $\varepsilon > 0$ with $\varepsilon \leq A(\lambda)+\lambda+\epsilon$. For $g_{f,\lambda} \in \mathcal{G}_\varepsilon(\lambda)$ we then have $\|g_{f,\lambda}\|_\infty \leq 4\Lambda(\lambda,\varepsilon)$ and $\mathbb{E}_\mathrm{P} g_{f,\lambda}^2 \leq 12c\lambda\Lambda^2(\lambda,\varepsilon)$ by Lemmas 3.1 and 3.2. Moreover, Lemma 3.3 shows

$$\mathbb{E}_{D\sim \mathrm{P}^n} e_i\bigl(\mathcal{G}_\varepsilon(\lambda), \|\cdot\|_{L_2(\mathrm{D})}\bigr) \leq c_p\, a\, \Lambda(\lambda,\varepsilon)\, i^{-\frac{1}{2p}}\,.$$

By symmetrization and [14, Theorem 7.16], which translates bounds on average entropy numbers into bounds on local Rademacher averages, we thus find a constant $C_p \geq 2c$ only depending on p and c such that

$$\begin{aligned}
\omega_{\mathrm{P},n}(\mathcal{G}(\lambda),\varepsilon) &:= \mathbb{E}_{D\sim \mathrm{P}^n} \sup_{\substack{g\in\mathcal{G}(\lambda)\\ \mathbb{E}_\mathrm{P} g \leq \varepsilon}} \bigl|\mathbb{E}_\mathrm{P} g - \mathbb{E}_\mathrm{D} g\bigr| \\
&= \mathbb{E}_{D\sim \mathrm{P}^n} \sup_{g\in\mathcal{G}_\varepsilon(\lambda)} \bigl|\mathbb{E}_\mathrm{P} g - \mathbb{E}_\mathrm{D} g\bigr| \\
&\leq C_p\, \Lambda(\lambda,\varepsilon) \max\bigl\{a^p \lambda^{\frac{1-p}{2}}\, n^{-\frac{1}{2}}, a^{\frac{2p}{1+p}} n^{-\frac{1}{1+p}}\bigr\}\,.
\end{aligned}$$

We now define $\varepsilon := A(\lambda)+\lambda+\epsilon$, which implies $\Lambda(\lambda,\varepsilon) = 2\varepsilon\lambda^{-1}$. For $K := 1024C_p^2$ and $\lambda^{1+p} n \geq a^{2p} K$ we hence obtain

$$\omega_{\mathrm{P},n}(\mathcal{G}(\lambda),\varepsilon) \leq 2\, C_p \varepsilon \max\bigl\{a^p \lambda^{-\frac{1+p}{2}}\, n^{-\frac{1}{2}}, a^{\frac{2p}{1+p}} \lambda^{-1} n^{-\frac{1}{1+p}}\bigr\} \leq \frac{\varepsilon}{16}\,.$$

We further write $\mathcal{F} := \mathcal{G}(\lambda)$ and

$$C\circ f := \lambda\|f\|_E + L\circ f\,.$$

For $\beta := 1$, $b := 4/\lambda$ and $B := 4\cdot\frac{A(\lambda)+\lambda+\epsilon}{\lambda}$ we then see that the supremum bound (6) of [15, Theorem 3.1] is satisfied. Moreover, the variance bound (7) of [15, Theorem

3.1] is satisfied for $v := \vartheta := 1$, $w := 3c$, and $W := 3c(A(\lambda) + \lambda + \epsilon)$. In addition, $\lambda n \geq K\tau$ and $C_p \geq 2c$ imply

$$\sqrt{\frac{2\tau(b\varepsilon^\beta + B)^\nu(w\varepsilon^\vartheta + W)}{n}} = \sqrt{\frac{24\tau c\lambda\Lambda^2(\lambda,\varepsilon)}{n}} \leq \sqrt{\frac{96c}{K}}\varepsilon \leq \frac{\varepsilon}{4}$$

and

$$\frac{2\tau\big(b\varepsilon^\beta + B\big)}{n} = \frac{8\tau\Lambda(\lambda,\varepsilon)}{n} = \frac{8\tau\varepsilon}{\lambda n} \leq \frac{\varepsilon}{128}\,.$$

Using these estimates together with a repetition of the proof of [15, Theorem 3.1] for $a := 1/2$ instead of $a = 0$, we further see that every $\varepsilon/2$-ARBA satisfies

$$\lambda\|f_{D,\lambda}\|_E + \mathcal{R}_{L,\mathrm{P}}(f_{D,\lambda}) - \mathcal{R}^*_{L,\mathrm{P}} < \lambda\|f_{\mathrm{P},\lambda}\|_E + \mathcal{R}_{L,\mathrm{P}}(f_{\mathrm{P},\lambda}) - \mathcal{R}^*_{L,\mathrm{P}} + \varepsilon$$

with probability P^n not smaller than $1 - e^{-\tau}$. Since we obviously have $\lambda/2 \leq \varepsilon/2$, we then obtain the assertion for $\epsilon \to 0$. ■

3.2 Proofs of the results related to Theorem 2.10

Proof of Lemma 2.5: The assertion immediately follows by combining [14, Lemma 2.23], [14, Lemma 3.12], and [14, Lemma 3.64]. ■

For the proof of Lemma 2.7 and Proposition 2.9 we need some preparations. To this end, we define, as in [14, Chapter 3], the inner risk of a loss function $L : Y \times \mathbb{R} \to [0,\infty)$ by

$$\mathcal{C}_{L,\eta}(t) := \int_Y L(y,t)\,d\mathrm{Q}(y) = \eta L(1,t) + (1-\eta)L(-1,t)\,, \qquad t \in \mathbb{R}\,,$$

where Q is a distribution on Y and $\eta := \mathrm{Q}(\{1\})$. Obviously, the L-risk of a function $f : X \to \mathbb{R}$ can then be computed by

$$\mathcal{R}_{L,\mathrm{P}}(f) = \int_X \mathcal{C}_{L,\eta(x)}(f(x))\,d\mathrm{P}_X\,. \tag{16}$$

Moreover, for $0 < \delta < 1/2$ we define $\eta^{(\delta)} := (1-\delta)\eta + \delta(1-\eta)$. A simple calculation then shows

$$\begin{aligned}
\mathcal{C}_{L_\delta,\eta}(t) &= \eta(1-\delta)L(1,t) + \eta\delta L(-1,t) \\
&\quad +(1-\eta)(1-\delta)L(-1,t) + (1-\eta)\delta L(1,t) \\
&= \eta^{(\delta)}L(1,t) + (1-\eta^{(\delta)})L(-1,t) \\
&= \mathcal{C}_{L,\eta^{(\delta)}}(t)
\end{aligned} \tag{17}$$

for all $t \in \mathbb{R}$. Obviously, if we define the minimal inner risk of a loss L by

$$\mathcal{C}^*_{L,\eta} := \inf_{t\in\mathbb{R}} \mathcal{C}_{L,\eta}(t)\,,$$

then Equation (17) yields $\mathcal{C}^*_{L_\delta,\eta} = \mathcal{C}^*_{L,\eta^{(\delta)}}$. Furthermore, we finally need the set

$$\mathcal{M}_{L,\eta}(0^+) := \left\{t \in \mathbb{R} : \mathcal{C}_{L,\eta}(t) = \mathcal{C}^*_{L,\eta}\right\},$$

which contains all global minimizers of $t \mapsto \mathcal{C}_{L,\eta}(t)$. Note that we always have $\mathcal{C}^*_{L,\eta} < \infty$, and hence the definition of $\mathcal{M}_{L,\eta}(0^+)$ coincides with that on page 53 of [14]. Moreover, our considerations above show $\mathcal{M}_{L_\delta,\eta}(0^+) = \mathcal{M}_{L,\eta^{(\delta)}}(0^+)$. The following lemma collects some useful properties of the latter set.

Lemma 3.4 *Let $L \neq 0$ be a convex, classification calibrated, and margin-based loss represented by $\varphi : \mathbb{R} \to [0,\infty)$. Moreover, we fix a $\delta \in \mathbb{R}$ with $0 < \delta < 1/2$. Then, for all $\eta \in [\delta, 1-\delta]$, the function*

$$t \mapsto \mathcal{C}_{L,\eta}(t)$$

has a global minimum, i.e., $\mathcal{M}_{L,\eta}(0^+) \neq \emptyset$. Writing

$$M_\delta := \inf \mathcal{M}_{L,1-\delta}(0^+) = \inf\left\{t \in \mathbb{R} : L_\delta(1,t) \leq L_\delta(1,s) \text{ for all } s \in \mathbb{R}\right\},$$

we further have $0 < M_\delta < \infty$ and

$$\mathcal{M}_{L,\eta}(0^+) \subset [-M_\delta, M_\delta], \qquad \eta \in (\delta, 1-\delta). \tag{18}$$

Moreover, if φ is strictly convex, $\mathcal{M}_{L,\eta}(0^+)$ contains exactly one element, denoted by t^*_η, for all $n \in [\delta, 1-\delta]$. In this case,

$$\eta \to t^*_\eta$$

is a monotonically increasing function on $[\delta, 1-\delta]$, and the restriction $\varphi_{|[-M_\delta, M_\delta]}$ of φ to $[-M_\delta, M_\delta]$ is strictly decreasing.

Proof: The convexity of the representing function φ of L implies $\lim_{t\to-\infty}\varphi(t) = \infty$ or $\lim_{t\to\infty}\varphi(t) = \infty$. From this we conclude that $\lim_{t\to\pm\infty}\mathcal{C}_{L,\eta}(t) = \infty$, and hence the convexity of $t \mapsto \mathcal{C}_{L,\eta}(t)$ shows that this function has a global minimum.

To show the second assertion, we first observe that $M_\delta > 0$ by the classification calibration of L and [14, Lemma 3.33]. Moreover, [14, Lemma 8.31] yields

$$\sup \mathcal{M}_{L,\eta}(0^+) \leq \inf \mathcal{M}_{L,1-\delta}(0^+) = M_\delta, \qquad \eta < 1-\delta$$

and

$$\inf \mathcal{M}_{L,\eta}(0^+) = -\sup \mathcal{M}_{L,1-\eta}(0^+) \geq -\inf \mathcal{M}_{L,1-\delta}(0^+) = -M_\delta$$

for all $\eta > \delta$.

For the proof of the last assertion, we first observe that the strict convexity of φ implies that $t \mapsto \mathcal{C}_{L,\eta}(t)$ is strictly convex, and hence this function has indeed a unique global minimizer. The monotonicity of $\eta \to t^*_\eta$ then follows by another application of [14, Lemma 8.31]. Finally, if φ is decreasing, the last assertion is trivial. On the other hand,

if φ is not decreasing, [14, Lemma 8.37] shows that $t_0 := \inf\{t \in \mathbb{R} : 0 \in \partial\varphi(t)\}$, where $\partial\varphi(t)$ denotes the subdifferential of φ at t, satisfies $0 < t_0 < \infty$ and $0 \in \partial\varphi(t_0)$. Consequently, we have $t_0 = \inf \mathcal{M}_{L,1}(0^+)$, and hence we obtain

$$M_\delta \leq \sup \mathcal{M}_{L,1-\delta}(0^+) \leq \inf \mathcal{M}_{L,1}(0^+) = t_0$$

by yet another application of [14, Lemma 8.31]. Moreover, by the convexity of φ, we see that φ is strictly decreasing on $(-\infty, t_0]$, and hence the last assertion follows. ■

Proof of Lemma 2.7: *i).* Trivial.

ii). Since $L_\delta(1,t) = \mathcal{C}_{L,1-\delta}(t)$ and $L_\delta(-1,t) = \mathcal{C}_{L,\delta}(t)$ for all $t \in \mathbb{R}$, we see by Lemma 3.4 that the functions $t \mapsto L_\delta(1,t)$ and $t \mapsto L_\delta(-1,t)$ have global minima at M_δ and $-M_\delta$, respectively. Using [14, Lemma 2.23], we then obtain the assertion.

iii). This is an immediate consequence of (17). ■

Proof of Proposition 2.9: Following Lemma 3.4 we denote the unique minimizer of $t \mapsto \mathcal{C}_{L,\eta}(t)$ by t^*_η. By (17) and $\eta^{(\delta)} \in [\delta, 1-\delta]$ it is easy to see that $t^*_{\eta^{(\delta)}}$ is the unique minimizer of $t \mapsto \mathcal{C}_{L_\delta,\eta}(t)$. For $\eta \in [0,1]$ and $t \in \mathbb{R}$ we now define

$$\begin{aligned} M(\eta,t) &:= \eta\big(L_\delta(1,t) - L_\delta(1,t^*_{\eta^{(\delta)}})\big)^2 + (1-\eta)\big(L_\delta(-1,t) - L_\delta(-1,t^*_{\eta^{(\delta)}})\big)^2 \\ E(\eta,t) &:= \eta\big(L_\delta(1,t) - L_\delta(1,t^*_{\eta^{(\delta)}})\big) + (1-\eta)\big(L_\delta(-1,t) - L_\delta(-1,t^*_{\eta^{(\delta)}})\big)\,. \end{aligned}$$

Obviously, it then suffices to show

$$M(\eta,t) \leq \big(\varphi(M_\delta) + \varphi(-M_\delta) + C_L(\delta)\big)E(\eta,t) \tag{19}$$

for all $\eta \in [0,1]$ and $t \in [-M_\delta, M_\delta]$. To this end, we further define

$$\begin{aligned} N(\eta,t) &:= \eta\big(\varphi(t) - \varphi(t^*_\eta)\big)^2 + (1-\eta)\big(\varphi(-t) - \varphi(-t^*_\eta)\big)^2 \\ D(\eta,t) &:= \eta\big(\varphi(t) - \varphi(t^*_\eta)\big) + (1-\eta)\big(\varphi(-t) - \varphi(-t^*_\eta)\big) \end{aligned}$$

for $\eta \in [\delta, 1-\delta]$ and $t \in [-M_\delta, M_\delta]$. Since $\mathcal{C}_{L,\eta}(t) = \eta\varphi(t) + (1-\eta)\varphi(-t)$, $t \in \mathbb{R}$, the minimizer t^*_η satisfies

$$\eta\varphi'(t^*_\eta) = (1-\eta)\varphi'(-t^*_\eta) \tag{20}$$

for all $\eta \in [\delta, 1-\delta]$. As in the proof of [4, Lemma 19], we thus obtain

$$\frac{\partial D}{\partial \eta}(\eta,t) = \big(\varphi(t) - \varphi(t^*_\eta)\big) - \big(\varphi(-t) - \varphi(-t^*_\eta)\big)$$

and

$$\begin{aligned}
&\frac{\partial N}{\partial \eta}(\eta,t)\\
&= \big(\varphi(t)-\varphi(t_\eta^*)\big)^2-\big(\varphi(-t)-\varphi(-t_\eta^*)\big)^2\\
&\quad -2\eta\big(\varphi(t)-\varphi(t_\eta^*)\big)\varphi'(t_\eta^*)\frac{\partial t_\eta^*}{\partial \eta}+2(1-\eta)\big(\varphi(-t)-\varphi(-t_\eta^*)\big)\varphi'(-t_\eta^*)\frac{\partial t_\eta^*}{\partial \eta}\\
&= \Big(\big(\varphi(t)-\varphi(t_\eta^*)\big)-\big(\varphi(-t)-\varphi(-t_\eta^*)\big)\Big)\Big(\big(\varphi(t)-\varphi(t_\eta^*)\big)+\big(\varphi(-t)-\varphi(-t_\eta^*)\big)\Big)\\
&\quad -2\eta\varphi'(t_\eta^*)\Big(\big(\varphi(t)-\varphi(t_\eta^*)\big)-\big(\varphi(-t)-\varphi(-t_\eta^*)\big)\Big)\frac{\partial t_\eta^*}{\partial \eta}\\
&= \left(\big(\varphi(t)-\varphi(t_\eta^*)\big)+\big(\varphi(-t)-\varphi(-t_\eta^*)\big)-2\eta\varphi'(t_\eta^*)\frac{\partial t_\eta^*}{\partial \eta}\right)\frac{\partial D}{\partial \eta}(\eta,t)\,,
\end{aligned}$$

where we used (20) to obtain the second equality. We now define

$$C_\delta := \sup\left\{-\varphi(t_\eta^*)-\varphi(-t_\eta^*)-2\eta\varphi'(t_\eta^*)\frac{\partial t_\eta^*}{\partial \eta} : \eta\in[\delta,1-\delta]\right\},$$

and observe by the last assertion of Lemma 3.4 that $\frac{\partial D}{\partial \eta}(\eta,t)\geq 0$ if and only if $t\leq t_\eta^*$. Consequently, we find

$$\frac{\partial N}{\partial \eta}(\eta,t)\leq\big(\varphi(t)+\varphi(-t)+\max\{0,C_\delta\}\big)\frac{\partial D}{\partial \eta}(\eta,t)$$

for all $\eta\in[\delta,1-\delta]$ and $t\in[-M_\delta,M_\delta]$ satisfying $t\leq t_\eta^*$. Analogously, we obtain the inverse inequality for $t\geq t_\eta^*$. Since $N(\eta,t_\eta^*)=D(\eta,t_\eta^*)=0$ for all $\eta\in[0,1]$, the fundamental theorem of calculus thus shows

$$N(\eta,t)\leq\big(\varphi(t)+\varphi(-t)+\max\{0,C_\delta\}\big)\,D(\eta,t) \tag{21}$$

for all $\eta\in[\delta,1-\delta]$ and $t\in[-M_\delta,M_\delta]$. In order to estimate C_δ, we now observe that

$$\frac{\partial t_\eta^*}{\partial \eta}=-\frac{\varphi'(t_\eta^*)+\varphi'(-t_\eta^*)}{\eta\varphi''(t_\eta^*)+(1-\eta)\varphi''(-t_\eta^*)}\,,$$

and hence (20) yields

$$\begin{aligned}
-2\eta\varphi'(t_\eta^*)\frac{\partial t_\eta^*}{\partial \eta} &= \big(\varphi'(t_\eta^*)+\varphi'(-t_\eta^*)\big)\cdot\frac{\eta\varphi'(t_\eta^*)+(1-\eta)\varphi'(-t_\eta^*)}{\eta\varphi''(t_\eta^*)+(1-\eta)\varphi''(-t_\eta^*)}\\
&= \big(\varphi'(t_\eta^*)+\varphi'(-t_\eta^*)\big)\cdot\frac{2}{\frac{\varphi''(t_\eta^*)}{\varphi'(t_\eta^*)}+\frac{\varphi''(-t_\eta^*)}{\varphi'(-t_\eta^*)}}\,.
\end{aligned}$$

From this we conclude

$$C_\delta=\sup\left\{\frac{2\varphi'(t_\eta^*)\varphi'(-t_\eta^*)(\varphi'(t_\eta^*)+\varphi'(-t_\eta^*))}{\varphi'(-t_\eta^*)\varphi''(t_\eta^*)+\varphi'(t_\eta^*)\varphi''(-t_\eta^*)}-\varphi(t_\eta^*)-\varphi(-t_\eta^*) : \eta\in[\delta,1-\delta]\right\}.$$

Furthermore, the monotonicity of φ on $[-M_\delta, M_\delta]$, see Lemma 3.4, yields

$$\big(\varphi(t) - \varphi(t^*_{\eta^{(\delta)}})\big)\big(\varphi(-t) - \varphi(-t^*_{\eta^{(\delta)}})\big) \leq 0$$

for all $\eta \in [0,1]$ and all $t \in [-M_\delta, M_\delta]$, and hence we have

$$\begin{aligned}
M(\eta,t) &= \eta\big((1-\delta)(\varphi(t) - \varphi(t^*_{\eta^{(\delta)}})) + \delta(\varphi(-t) - \varphi(-t^*_{\eta^{(\delta)}}))\big)^2 \\
&\quad +(1-\eta)\big((1-\delta)(\varphi(-t) - \varphi(-t^*_{\eta^{(\delta)}})) + \delta(\varphi(t) - \varphi(t^*_{\eta^{(\delta)}}))\big)^2 \\
&= \eta(1-\delta)^2\big(\varphi(t) - \varphi(t^*_{\eta^{(\delta)}})\big)^2 \\
&\quad +\eta\delta^2\big(\varphi(-t) - \varphi(-t^*_{\eta^{(\delta)}})\big)^2 \\
&\quad +(1-\eta)\delta^2\big(\varphi(t) - \varphi(t^*_{\eta^{(\delta)}})\big)^2 \\
&\quad +(1-\eta)(1-\delta)^2\big(\varphi(-t) - \varphi(-t^*_{\eta^{(\delta)}})\big)^2 \\
&\quad +2\delta(1-\delta)\big(\varphi(t) - \varphi(t^*_{\eta^{(\delta)}})\big)\big(\varphi(-t) - \varphi(-t^*_{\eta^{(\delta)}})\big) \\
&\leq \eta(1-\delta)\big(\varphi(t) - \varphi(t^*_{\eta^{(\delta)}})\big)^2 \\
&\quad +\eta\delta\big(\varphi(-t) - \varphi(-t^*_{\eta^{(\delta)}})\big)^2 \\
&\quad +(1-\eta)\delta\big(\varphi(t) - \varphi(t^*_{\eta^{(\delta)}})\big)^2 \\
&\quad +(1-\eta)(1-\delta)\big(\varphi(-t) - \varphi(-t^*_{\eta^{(\delta)}})\big)^2 \\
&= \eta^{(\delta)}\big(\varphi(t) - \varphi(t^*_{\eta^{(\delta)}})\big)^2 + (1-\eta^{(\delta)})\big(\varphi(-t) - \varphi(-t^*_{\eta^{(\delta)}})\big)^2 \\
&= N(\eta^{(\delta)}, t)
\end{aligned}$$

where in the inequality we used $\delta^2 \leq \delta$ and $(1-\delta)^2 \leq 1-\delta$. Moreover, by (17) we have

$$D(\eta^{(\delta)}, t) = \mathcal{C}_{L,\eta^{(\delta)}}(t) - \mathcal{C}^*_{L,\eta^{(\delta)}} = \mathcal{C}_{L_\delta,\eta}(t) - \mathcal{C}^*_{L_\delta,\eta} = E(\eta,t)$$

for all $\eta \in [0,1]$ and $t \in \mathbb{R}$. In addition, we have

$$\varphi(t) + \varphi(-t) = 2\mathcal{C}_{L,1/2}(t) \leq 2\mathcal{C}_{L,1/2}(M_\delta) = \varphi(M_\delta) + \varphi(-M_\delta)$$

for all $t \in [-M_\delta, M_\delta]$ by the convexity and symmetry of $t \mapsto \mathcal{C}_{L,1/2}(t)$. Combining these considerations with (21) and both $\eta^{(\delta)} \in [\delta, 1-\delta]$ and $t^*_{\eta^{(\delta)}} \in [-M_\delta, M_\delta]$, we then obtain (19). ∎

Proof of Theorem 2.10: Our goal is to apply [14, Theorem 7.20]. To this end we define $\Upsilon : E \to [0,\infty)$ by $\Upsilon(f) := \lambda\|f\|_E$, $f \in E$. By Lemma 3.4 we recall that φ is strictly decreasing on the interval $[-M_\delta, M_\delta]$, and hence we easily find

$$L_\delta(y,t) \leq \varphi(-M_\delta)\,, \qquad\qquad y = \pm 1,\, t \in [-M_\delta, M_\delta]\,,$$

i.e., the supremum bound (7.35) in [14] is satisfied for $B := \varphi(-M_\delta)$. Moreover, Proposition 2.9 shows that the variance bound (7.36) in [14] is satisfied for $\vartheta := 1$ and $V := \varphi(M_\delta) + \varphi(-M_\delta) + C_L(\delta)$. In addition, we obviously have $V \geq B^{2-\vartheta}$. In the

following, we pick a function $f_0 \in E$ with $\lambda\|f_0\|_E + \mathcal{R}_{L_\delta,\mathrm{P}}(f_0) - \mathcal{R}^*_{L_\delta,\mathrm{P}} \leq 2A(\lambda)$. The assumptions on L then yield

$$\|L_\delta \circ f_0\|_\infty \leq 1 + \|f_0\|_\infty \leq 1 + \frac{2A(\lambda)}{\lambda}\,.$$

In the following, we thus set $B_0 := B + \frac{2A(\lambda)}{\lambda}$. Last but not least, the definitions (7.32) – (7.34) in [14] become

$$\begin{aligned} r^* &:= \inf_{f\in E} \lambda\|f\|_E + \mathcal{R}_{L_\delta,\mathrm{P}}(\widehat{f}\,) - \mathcal{R}^*_{L_\delta,\mathrm{P}} \\ \mathcal{F}_r &:= \{f \in E : \lambda\|f\|_E + \mathcal{R}_{L_\delta,\mathrm{P}}(\widehat{f}\,) - \mathcal{R}^*_{L_\delta,\mathrm{P}} \leq r\}\,, \\ \mathcal{H}_r &:= \{L_\delta \circ \widehat{f} - L_\delta \circ f^*_{L_\delta,\mathrm{P}} : f \in \mathcal{F}_r\}\,, \end{aligned}$$

where the latter two sets are only defined for $r > r^*$. Now observe that for $f \in \mathcal{F}_r$ we have $\lambda\|f\|_E \leq r$, and hence we conclude that $\mathcal{F}_r \subset \frac{r}{\lambda} B_E$. Since L_δ is Lipschitz continuous with $|L_\delta|_1 \leq |L|_1 \leq 1$, we thus find

$$\mathbb{E}_{D\sim \mathrm{P}^n} e_i(\mathcal{H}_r, L_2(\mathrm{D})) \leq \mathbb{E}_{D_X\sim \mathrm{P}_X^n} e_i(\mathcal{F}_r, L_2(\mathrm{D}_X)) \leq 2r\lambda^{-1} a\, i^{-\frac{1}{2p}}\,.$$

Moreover, for $f \in \mathcal{F}_r$, we have $\mathbb{E}_\mathrm{P}(L \circ \widehat{f} - L \circ f^*_{L,\mathrm{P}})^2 \leq Vr$, and consequently [14, Theorem 7.16] shows that the Rademacher average in (7.37) of [14] is bounded by the function

$$\varphi_n(r) := c_L(\delta,p) \max\left\{ a^p r^{\frac{1+p}{2}} \lambda^{-p} n^{-\frac{1}{2}}, a^{\frac{2p}{1+p}} r^{\frac{2p}{1+p}} \lambda^{-\frac{2p}{1+p}} n^{-\frac{1}{1+p}} \right\},$$

where $c_L(\delta,p) \geq 1$ is a constant only depending on L, δ, and p. Obviously, this function does in general not satisfy the condition $\varphi_n(4r) \leq 2\varphi_n(r)$, $r \geq r^*$, required in [14, Theorem 7.20]. However, it satisfies $\varphi_n(4r) \leq 4^{\frac{1+p}{2}}\varphi_n(r)$, $r \geq r^*$, and since $\frac{1+p}{2} < 1$ is all we need for the peeling argument [14, Theorem 7.7] employed in the proof of [14, Theorem 7.20], Condition (7.38) in [14] only changes by a constant c_p in front of $30\varphi_n(r)$. Consequently, Condition (7.38) reduces to

$$r > \tilde{c}_L(\delta,p)\Big(\frac{a^{2p}}{\lambda^{2p} n}\Big)^{\frac{1}{1-p}} + \frac{77V\tau}{n} + \frac{10\tau A(\lambda)}{\lambda n} + A(\lambda)$$

From this we easily obtain the assertion by [14, Theorem 7.20]. ∎

References

[1] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.*, 101:138–156, 2006.

[2] P. L. Bartlett and M. Traskin. AdaBoost is consistent. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 105–112, Cambridge, MA, 2007. MIT Press.

[3] G. Blanchard, O. Bousquet, and P. Massart. Statistical performance of support vector machines. *Ann. Statist.*, 36:489–531, 2008.

[4] G. Blanchard, G. Lugosi, and N. Vayatis. On the rate of convergence of regularized boosting classifiers. *J. Mach. Learn. Res.*, 4:861–894, 2003.

[5] B. Carl, I. Kyrezi, and A. Pajor. Metric entropy of convex hulls in Banach spaces. *J. London Math. Soc.*, 60:871–896, 1999.

[6] B. Carl and I. Stephani. *Entropy, Compactness and the Approximation of Operators*. Cambridge University Press, Cambridge, 1990.

[7] R. M. Dudley. Universal Donsker classes and metric entropy. *Ann. Probab.*, 15:1306–1326, 1987.

[8] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.

[9] I. Kyrezi. On the entropy of the convex hull of finite sets. *Proc. Amer. Math. Soc.*, 128:2393–2403, 2000.

[10] G. Lugosi and N. Vayatis. On the Bayes-risk consistency of regularized boosting methods. *Ann. Statist.*, 32:30–55, 2004.

[11] R. Meir and G. Rätsch. An introduction to boosting and leveraging. In S. Mendelson and A. J. Smola, editors, *Advanced Lectures on Machine Learning*, pages 119–184. Springer, Berlin, 2003.

[12] I. Steinwart. Entropy of $C(K)$-valued operators. *J. Approx. Theory*, 103:302–328, 2000.

[13] I. Steinwart. Some bounds on random entropy numbers with an application to support vector machines. Technical Report LA-UR-08-3251, Los Alamos National Laboratory, 2008.

[14] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, New York, 2008.

[15] I. Steinwart, D. Hush, and C. Scovel. A new concentration result for regularized risk minimizers. In E. Giné, V. Koltchinskii, W. Li, and J. Zinn, editors, *High Dimensional Probability IV*, pages 260–275. Institute of Mathematical Statistics, Beachwood, OH, 2006.

[16] A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32:135–166, 2004.

[17] Y. Yang. Minimax nonparametric classification—part I and II. *IEEE Trans. Inform. Theory*, 45:2271–2292, 1999.

[18] T. Zhang and B. Yu. Boosting with early stopping: Convergence and consistency. *Ann. Statist.*, 33:1538–1579, 2005.